Comparación de algoritmos detectores de puntos singulares para reconocimiento de objetos en vídeo quirúrgico.

I. García Barquero[1, 2], P. Sánchez-González[1,2], M. Luna Serrano[1,2], E.J. Gómez Aguilera[1,2]

[1] Grupo de Bioingeniería y Telemedicina, ETSI Telecomunicación, Universidad Politécnica de Madrid, Madrid, España, {igarcia, psanchez, mluna, egomez}@gbt.tfo.upm.es

[2] Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina, Zaragoza, España

Resumen

El análisis de vídeo laparoscópico ofrece nuevas posibilidades a la navegación quirúrgica al garantizar una incorporación mínima de tecnología en quirófano, evitando así alterar la ergonomía y los flujos de trabajo de las intervenciones. Una de sus principales ventajas es que puede servir como fuente de datos para reconstruir tridimensionalmente la escena laparoscópica, lo que permite dotar al cirujano de la sensación de profundidad perdida en este tipo de cirugía. En el presente trabajo de investigación se comparan dos detectores de puntos singulares, SIFT y SURF, para estimar cuál de los dos podría integrarse en un algoritmo de cálculo de coordenadas 3D, MonoSLAM, basado en la detección y el seguimiento de estos puntos singulares en los fotogramas del vídeo. Los resultados obtenidos posicionan a SURF como la mejor opción gracias a su rapidez y a su mayor capacidad de discriminación entre estructuras anatómicas e instrumental quirúrgico.

1. Introducción

La Cirugía de Mínima Invasión (CMI) ha crecido de manera espectacular en los últimos años y se ha introducido y consolidado en las principales especialidades quirúrgicas de la práctica clínica gracias a las múltiples ventajas que ofrece al paciente, entre las que destacan menor trauma tisular, menor morbilidad, menor estancia hospitalaria y recuperación más rápida. Sin embargo, los cirujanos han de afrontar nuevos retos como la visualización de la escena a través de monitores, el manejo de nuevos instrumentos y la pérdida de percepción de profundidad [1].

La información tridimensional de la escena quirúrgica ofrece al cirujano una localización precisa de la posición y de la profundidad a la que se encuentran los diferentes agentes que aparecen en la escena (instrumental y estructuras anatómicas) [2], lo que es de gran utilidad para los navegadores quirúrgicos. Así, se podría alertar al cirujano cuando un determinado instrumento se encuentre demasiado cerca de un tejido sensible, como capilares o venas que podrían producir hemorragias indeseables.

Los navegadores se basan, en gran medida, en la introducción en quirófano de sistemas de imagen (Tomografía computarizada (TC), Ultrasonido laparoscópico (UL)) y de tracking (mediante sensores mecánicos, electromagnéticos, acústicos u ópticos) [3][4], que localizan y hacen un seguimiento de los diferentes elementos que aparecen en la intervención. Su presencia no sólo supone ocupar un valioso espacio dentro de la sala de operaciones sino que su manejo modifica la ergonomía del quirófano e interrumpe ocasionalmente los flujos de trabajo de los cirujanos [3].

Los avances tecnológicos están permitiendo que los navegadores sean cada vez menos intrusivos, como los basados en el análisis de vídeo laparoscópico, ya que se minimiza la tecnología en la sala al ser el vídeo un elemento imprescindible en este tipo de cirugía [4]. Actualmente, el navegador hepático THEMIS [3] explota la información del vídeo endoscópico para ofrecer un sistema de tracking de instrumentos y órganos y dotar al cirujano de sensación de profundidad estimando la tridimensionalidad de la escena. Dicha sensación se implementa mediante una técnica de visión por ordenador, conocida como Shape from Shading, que utiliza información de iluminación y sombras de un fotograma [2][5]. Su principal inconveniente es que el mapa de profundidades que genera no es absoluto, siendo necesario combinarla con otras técnicas para completar la información. En la literatura se han encontrado referencias a otras metodologías que emplean señales de profundidad diferentes: la visión estéreo [6][7] y el movimiento [8][9].

Las técnicas *Shape from Stereo* [6][7] se basan en el modelo de visión estereoscópica por lo que requieren el uso de laparoscopios estéreo. Su principal inconveniente es que son instrumentos no aceptados por la comunidad clínica debido a su alto precio, haciendo que esta alternativa no sea realmente viable. Las técnicas *Shape from Motion* [8][9], permiten hallar las coordenadas 3D de los objetos a partir de secuencias de vídeo [10]. Ya se han aplicado con éxito en campos diferentes a la bioingeniería, como la robótica o la defensa [11].

Este trabajo de investigación se centra en una técnica *Shape from Motion* conocida como MonoSLAM. Es un algoritmo que reconstruye la trayectoria 3D de una cámara monocular que se mueve en un escenario a priori desconocido y, a su vez, genera en tiempo real un mapa de profundidades de dicho escenario gracias a puntos de referencia naturales detectados en las imágenes que captura la propia cámara monoscópica [11].

Uno de los principales retos que conlleva MonoSLAM es la detección, identificación y seguimiento de puntos singulares de la escena a lo largo de la intervención quirúrgica. El objetivo del presente trabajo de investigación es analizar dos de los algoritmos detectores de puntos singulares más utilizados en la literatura, SIFT y SURF, adaptados a las imágenes de laparoscopia. Se valoran las bondades y desventajas de cada uno, buscando cuál de ellos presenta un comportamiento más estable cuando, en los fotogramas del vídeo laparoscópico, se producen cambios en las condiciones de iluminación, escala o ángulo de visión.

Además, MonoSLAM requiere una posterior etapa de emparejamiento entre los puntos detectados. El algoritmo escogido es el de "Vecino más cercano", que se basa en el cálculo de distancias euclídeas. De esta forma, se evalúa también la fiabilidad de los emparejamientos entre los puntos singulares detectados por SIFT y SURF, aspecto clave para decantarnos por uno de ambos algoritmos.

2. Material y métodos

Los datos a procesar proceden de tres secuencias de vídeo laparoscópico, en los que aparecen tres, uno y ningún instrumentos laparoscópicos. Sus principales características se resumen en la Tabla 1.

	3 instrum.	1 instrum.	Sin instrum.
Tasa (fps)	30	30	30
Duración (s)	85	4	16
Nº fotogramas	2547	124	510
Tamaño fotogramas	720 x 576	512 x 384	720 x 576
Formato	.jpg	.jpg	.jpg
Relación de aspecto	PAL 4:3	Square pixel	PAL 4:3

Tabla 1. *Características de las tres secuencias de vídeo*

Para su descomposición en fotogramas se ha utilizado el programa TMPGEnc Mastering Works 5 (Pegasys Inc., Tokyo, Japón) y para su análisis, la herramienta matemática MATLAB versión R2011a (Mathworks, Natick, MA, U.S.A). Para el procesado se ha empleado un procesador Intel Core 2 Duo P8600, a 2.40 GHz con 4 GB de memoria RAM.

Los métodos utilizados en el trabajo han sido los algoritmos detectores SIFT y SURF, la técnica de reconstrucción 3D MonoSLAM y el algoritmo de emparejamientos "Vecino más cercano", que se describen en los siguientes sub-apartados.

2.1. Metodología MonoSLAM

MonoSLAM (Monocular Simultaneous Localization And Mapping) se basa en la metodología SLAM, una plataforma con sensores en movimiento que reconstruye su entorno de forma tridimensional a la vez que estima su propia localización. Por tanto, MonoSLAM es un caso particular de SLAM que extrae los datos del exterior mediante cámaras monoculares. Se trata de una metodología "top-down" que crea un mapa probabilístico a partir de un conjunto de características naturales del entorno, representando en cada instante de tiempo la posición actual de la cámara y de todas las características de interés. Su fase crítica es la detección de los puntos de referencia, que en un principio se hallan mediante el operador de Shi y Tomasi [11]. Este operador será sustituido por SIFT o SURF, una vez se hayan evaluado las bondades y desventajas de cada algoritmo, con el objetivo de adaptar esta metodología al caso concreto de cirugía laparoscópica.

2.2. Algoritmo SIFT

El algoritmo SIFT [12] (Scale-Invariant Feature Transform) transforma los puntos que detecta sobre la imagen en coordenadas invariantes a la escala, la rotación y, parcialmente, a la iluminación y al punto de vista 3D. Cada punto de interés detectado se describe con un vector descriptor que recoge su posición, escala y orientación (generalmente de tamaño 128).

Este algoritmo consta de cuatro etapas: generación del espacio escala (Figura 1) mediante funciones diferencia de gaussianas (DoGs); localización de puntos invariantes (máximos y mínimos del espacio escala); asignación de orientación a los puntos (en base a direcciones locales de los gradientes); y generación del descriptor del punto.

2.3. Algoritmo SURF

El algoritmo SURF [13] (Speed Up Robust Feature) consta de cuatro etapas similares al algoritmo anterior. Su propósito es la extracción de puntos invariantes en imágenes. La principal diferencia radica en que SURF es computacionalmente más rápido, pretendiendo no sacrificar rendimiento. El vector descriptor que crea generalmente tiene un tamaño de 64.

Las etapas de este algoritmo son: generación de imagen integral para agilizar los cálculos; creación del espacio escala (Figura 1) mediante aproximaciones a la segunda derivada de la gaussiana; localización de puntos invariantes (máximos y mínimos del espacio escala); asignación de orientación (con wavelets Haar); y generación del descriptor.

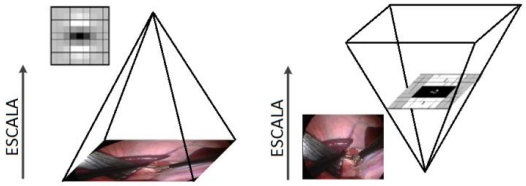

Figura 1. *Espacio escala mediante SIFT (izq) y SURF (dcha)*

2.4. Algoritmo "Vecino más cercano"

El algoritmo de emparejamiento de puntos singulares utilizado es el de "Vecino más cercano", que se basa en el cálculo de la distancia euclídea entre los vectores descriptores de los puntos singulares .

Para aumentar la fiabilidad de los emparejamientos, además, se impone un umbral entre las distancias del vecino más cercano y del segundo más cercano. Con ello, se garantiza que el vecino más cercano sea un punto más distintivo, reduciendo en un 90% los falsos positivos [12].

2.5. Procedimiento experimental

Para evaluar las bondades y desventajas de los algoritmos detectores se presentan dos conjuntos de pruebas: aquellas destinadas a valorar la calidad en la detección de los puntos singulares y aquellas que evalúan la fiabilidad del algoritmo de emparejamiento. Parte de dichas pruebas requiere la elaboración de una batería de imágenes en las que se van modificado de forma progresiva la iluminación, el ángulo de visión y la escala de un fotograma de vídeo laparoscópico (Tabla 2). El objetivo es estimar la estabilidad en la detección y distribución de los puntos singulares de los algoritmos cuando ocurren este tipo de fenómenos durante las intervenciones.

Cambio	Leve	Moderado	Notable
Escala	Sin viñeteo	Inst. y órganos centro	Puntas de inst. y órganos centro
Iluminación	75%	50%	25%
Ángulo	Rotación leve	Rotación moderada	Rotación moderada y desplazamiento leve

Tabla 2. *Batería de fotogramas con simulación de cambios*

Para evaluar el algoritmo de emparejamiento se utilizan dos parámetros: el rendimiento (ratio entre el número de emparejamientos y el número medio de puntos detectado en los fotogramas) y el valor predictivo positivo (relación

entre el número de emparejamientos correctos y el número total de emparejamientos).

3. Resultados y Discusión

A partir de los resultados obtenidos de la comparación entre los algoritmos se selecciona el mejor candidato a ser integrado en MonoSLAM para poder reconstruir tridimensionalmente la escena laparoscópica.

3.1. Evaluación de la calidad en la detección de puntos de interés

En primer lugar, se analizan la cantidad y calidad de los puntos detectados por ambos algoritmos. Como los fotogramas de las secuencias de vídeo no son del mismo tamaño, no se puede establecer la comparación en base al número de puntos detectados (los fotogramas más grandes tienen más puntos). Como alternativa, se emplea el número de píxeles por punto, que indica cada cuántos píxeles del fotograma se localizaría un punto de interés si éstos estuvieran repartidos de forma homogénea por la escena. En la Tabla 3 se detalla el número de píxeles por punto de los fotogramas de cada vídeo, en media y de forma máxima y mínima. En todos los casos el número de píxeles por punto en SIFT es inferior al de SURF, lo que indica que se detectan en total más puntos (la razón es que SIFT puede detectar varios puntos en una localización que se diferencian en la escala y la orientación). Sin embargo, no por ofrecer más puntos singulares se trata de un mejor detector sino que habrá que analizar posteriormente si existe una buena fiabilidad en los emparejamientos, pues la cercanía entre ellos puede llevar a confusión en la etapa de correspondencias.

	SIFT			SURF		
Vídeo	Media	Máx	Mín	Media	Máx	Mín
3 inst.	476.7	1685.9	280.6	989.8	3190.2	590.8
1 inst.	370.2	454.1	310.1	762.0	840.2	685.0
sin inst.	1260.5	7975.4	476.1	1576.9	6912	711.4

Tabla 3. *Número de píxeles por punto para los diferentes vídeos*

Por otra parte, se analiza cómo afecta la presencia o ausencia de instrumental a la distribución de los puntos detectados por ambos algoritmos. La tendencia general es ubicar los puntos de interés sobre las zonas metálicas de los instrumentos y en los brillos de los órganos. Cuando no hay instrumentos, el número de píxeles por punto aumenta drásticamente en SIFT (2.64 veces respecto al vídeo de 3 instrumentos) y de forma menos notable en SURF (1.59 veces), lo que indica que SIFT localiza la mayoría de sus puntos singulares sobre el instrumental y, por tanto, distingue peor las estructuras anatómicas. Además, la distribución de los puntos de SURF sobre la escena es más homogénea que la de SIFT, que tiende a concentrarlos en los bordes e intersecciones de las estructuras. Esto supone una ventaja para el primero pues lo ideal es que los puntos estén suficientemente separados entre sí para evitar que el posterior algoritmo de emparejamiento establezca correspondencias erróneas.

Otro hecho destacable de estos resultados radica en las grandes diferencias entre el máximo y mínimo número de píxeles por punto en los vídeos de tres y ningún instrumentos.

Cuando en los fotogramas sólo aparecen estructuras anatómicas, esta notable diferencia se debe a los cambios de intensidad que aparecen entre los diferentes órganos y tejidos que conforman la escena. Ambos algoritmos encuentran puntos que presentan grandes variaciones de gradiente, por lo que se localizan más puntos de interés sobre las zonas de los fotogramas en las que existen cambios importantes de intensidad, como las áreas próximas a venas o a tejidos con surcos (Figura 2).

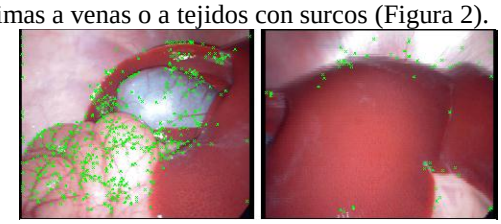

Figura 2. *Máximo (izq) y mínimo (dcha) número de puntos detectados por SURF en fotogramas con estructuras anatómicas.*

Al comparar los fotogramas para los que se consiguen el máximo y el mínimo número de píxeles en el vídeo de tres instrumentos, apreciamos una drástica alteración en la iluminación. Ninguno de los dos algoritmos tiene un buen comportamiento cuando ocurre este fenómeno y es que la reducción de puntos es del 83.35% en SIFT y del 81.48% en SURF. Cuando se produce un cambio en la iluminación de la escena se modifican los valores de intensidad de la imagen haciendo que las variaciones locales sean menores; por tanto, al empobrecer la iluminación la diferencia entre los gradientes son menores y, consecuentemente, el número de puntos detectados disminuye. Estos cambios de iluminación son frecuentes en cirugía laparoscópica, sobre todo cuando se practican ablaciones y como resultado se emborrona la escena.

Este fenómeno se corrobora mediante la batería de imágenes de prueba en las que se simula un descenso progresivo de la iluminación. Al reducir al 75%, al 50% y al 25% la iluminación del fotograma original, se obtiene un número de puntos de interés cada vez menor, perdurando sólo aquellos localizados en los brillos.

En esta línea, se simulan otros dos efectos que pueden ocurrir en cirugía laparoscópica: cambios en el ángulo de visión y en el zoom de la cámara. En el primer caso se obtienen resultados prometedores con ambos algoritmos, aunque SIFT es más repetitivo en la distribución general de los puntos y en la de los puntos aislados. Ante cambios de escala (Figura 3), se comprueba que la distribución general de los puntos sigue la tendencia habitual pero SIFT es más preciso a la hora de repetir puntos aislados sobre los fotogramas modificados.

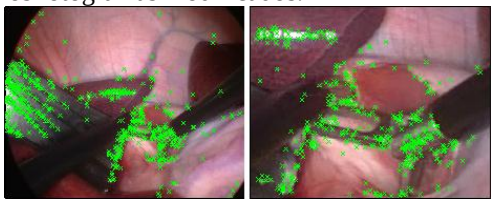

Figura 3. *Puntos detectados por SIFT ante cambios en la escala de la imagen*

Dicho comportamiento se debe a la mayor robustez de los vectores descriptores de SIFT, que emplean más dimensiones para caracterizar cada punto por lo que son más precisos a la hora de repetir las diferentes

localizaciones. Dicha robustez supone, sin embargo, una mayor carga computacional y, consecuentemente, un mayor tiempo de ejecución.

3.2. Evaluación de la fiabilidad en los emparejamientos

La elección del algoritmo "Vecino más cercano" se debe, principalmente, a su sencillez y bajo coste computacional [12]; lo que es una característica deseable al integrar todo el conjunto de técnicas de detección y emparejamiento en un navegador quirúrgico.

En la Tabla 4 se resumen, empleando diferentes umbrales entre las distancias de los descriptores del vecino más cercano y del segundo más cercano, el rendimiento y el VPP de los algoritmos.

	SIFT			SURF		
Umbral	0.9	0.8	**0.75**	0.9	0.8	**0.74**
Rend. (%).	64	49.88	44.65	68.95	57.25	53.43
VPP (%)	95.72	98.85	100	96.67	99.55	100

Tabla 4. *Rendimiento y VPP para diferentes umbrales*

Los resultados de SURF son más prometedores que los de SIFT al presentar menos falsos positivos para todos los umbrales y mejores rendimientos. Esta diferencia se debe a que el número de puntos detectados por SIFT es muy superior al de SURF, por lo que las correspondencias erróneas son más probables.

Por analogía con las pruebas anteriores, se utiliza la batería de fotogramas en los que se han simulado cambios de iluminación, ángulo de visión y escala para evaluar la fiabilidad de los emparejamientos. En la Tabla 5 se recogen los resultados obtenidos para los fotogramas en los que el cambio es moderado (ver Tabla 2). Los vídeos laparoscópicos no sufren una gran variabilidad por lo que las simulaciones moderadas son las más realistas.

	SIFT			SURF		
	Ilu	Ang.	Zoom	Ilu.	Ang.	Zoom
Rendimiento (%)	48.37	65.91	43.38	59.87	45.05	25.10
VPP (%)	95.69	99.26	96.77	84.78	100	100

Tabla 5. *Rendimiento y VPP para cambios en la escena*

Ante cambios en la iluminación, ninguno de los algoritmos ofrece un buen comportamiento, lo que corrobora la hipótesis inicial de no invariancia a este fenómeno. Este tipo de cambios son habituales en cirugía laparoscópica, por lo que actualmente se está trabajando en nuevos algoritmos que puedan ofrecer invariancia a estas alteraciones.

En cuanto a cambios en el ángulo de visión, la aparición de falsos positivos no es destacable, permitiendo que los VPPs sean cercanos al 100% (en ambos casos). Sin embargo, el rendimiento de SURF es más de un 20% inferior al de SIFT pues, como comprobamos en pruebas anteriores, SIFT genera vectores descriptores más robustos, lo que favorece la localización de los puntos incluso cuando el ángulo de visión ha variado.

Finalmente, al cambiar la escala de la escena comprobamos que ninguno de los dos algoritmos tiene un buen comportamiento pues SIFT intenta emparejar puntos del fotograma original que no existen en el fotograma con zoom y SURF establece correspondencias erróneas entre instrumental y estructuras anatómicas, independientemente de los buenos resultados de VPP obtenidos.

4. Conclusiones

El algoritmo SURF, por la forma en que se ha implementado, es un algoritmo más rápido que detecta más homogéneamente los puntos de interés por la escena laparoscópica, lo que permite emparejamientos más fiables en condiciones normales. Aunque SIFT presenta un comportamiento más estable cuando se producen cambios en el ángulo de visión, SURF también responde apropiadamente a este fenómeno y, como ninguno de los dos se posiciona como el mejor ante otro tipo de alteraciones (de iluminación y de escala), podemos concluir que SURF es el algoritmo con el que se detectarán más eficientemente los puntos singulares que servirán como referencia para reconstruir tridimensionalmente la escena laparoscópica mediante MonoSLAM, ofreciendo al cirujano sensación de profundidad a través de un navegador quirúrgico.

Referencias

[1] M. García, C. Toribio. El futuro de la Cirugía Mínimamente Invasiva: Tendencias tecnológicas a Medio y Largo Plazo, *Informe de prospectiva tecnológica del OPTI y FENIN*, 2004.

[2] P. Sanchez. Análisis de vídeo laparoscópico para formación en Cirugía de Mínima Invasión y Cirugía Guiada por Imagen, Tesis doctoral ETSIT (UPM), 2011.

[3] P. Sánchez-González et al. THEMIS: sistema de navegación quirúrgica en cirugía laparoscópica del hígado. *Congreso Anual de la Sociedad Española de Ingeniería Biomédica (CASEIB 2011)*, Cáceres, 2011.

[4] P. Sánchez-González et al. Laparoscopic video analysis for training and image-guided surgery, *Minimally Invasive Therapy and Allied Technologies*, vol 20, pp. 311-320 2011(ISSN: 1364-5706).

[5] T. Morvan et al. Exploiting Shading Information for Endoscopic Augmentation, *Workshop on Image Guidance and Computer Assistance for Soft-Tissue Interventions (IGSTI'08),* New York, 2008, pp 56-64.

[6] D. Stoyanov et al. Real-Time Stereo Reconstruction in Robotically Assisted Minimally Invasive Surgery. *Proceedings of Medical Image Computing and Computer Assisted Intervention*, vol. 1, 2010, pp. 275-282.

[7] D. P. Noonan et al. Stereoscopic Fibroscope for Camera Motion and 3D Depth Recovery during Minimally Invasive Surgery, *IEEE international conference on Robotics and Automation (ICRA'09),* Kobe, 2009, pp 3274-4279 (ISBN: 978-1-4244-2788-8).

[8] T. Collins et al. Deformable Shape-From-Motion in Laparoscopy using a Rigid Sliding Window, *Proceeding of the Medical Image Understanding and Analysis Conference (MIUA'11)*, London, 2011.

[9] A. Malti et al. Template-Based Deformable Shape-from Motion from Registered Laparoscopic Images, *Medical Image Understanding and Analysis (MIUA'11)*, London, 2011

[10] Levente Hajder, "Shape and motion from video".

[11] A. J. Davison et al. MonoSLAM: Real-Time Single Camera SLAM. *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol 29, 2007, pp 1052-1067 (ISSN: 0162-8828)

[12] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoint. *International Journal of Computer Vision*, vol 60, 2004, pp 91-110 (ISSN: 0920-5691)

[13] H. Bay et al. Speed-Up Robust Features (SURF), *Computer Vision and Image Understanding*, vol 110, 2008, pp 346-359 (ISSN: 1077-3142)